作业六：CLIP 图像分类

一、实验目的

通过使用 CLIP 模型来处理多模态目标分类任务。

1. 通过 CLIP 模型进行图像和文本的共同学习，探索如何利用图像与文本之间的关系来进行目标分类。
2. 比较不同的方法（如线性探针、零-shot 分类、适配器微调）对多模态目标分类任务的影响，特别是在不同准确率度量（Top-1 和 Top-5 准确率）下的表现。
3. 探索如何通过微调和适配器来提高模型在多模态任务中的分类精度，尤其是提高 Top-1 准确率。

二、实验设置

1. 数据集：
    - 使用的数据集包含了不同类别的图像，每个类别下有多个图像样本。
2. 模型：
    - CLIP 模型：利用 CLIP 模型，首先通过预训练的模型提取图像和文本的特征，然后通过不同的方法进行分类。
    - 零-shot 分类：直接使用 CLIP 模型的图像特征与文本特征进行余弦相似度计算，进行分类。
    - 线性探针：在 CLIP 模型的基础上使用逻辑回归进行图像特征分类。
    - 适配器微调：在 CLIP 模型中加入适配器层，进行微调，以增强图像和文本特征之间的对齐。
3. 实验过程：
    - 使用线性探针和适配器微调方法对模型进行训练。对于零-shot 分类，直接使用 CLIP 模型的预训练参数，不进行微调。
    - 在多个数据集上评估模型的 Top-1 和 Top-5 准确率。
    - 对比不同方法在同一数据集上的性能表现。

三、实验结果

1. 线性探针：
    - Top-1 准确率：64.49%
    - Top-5 准确率：86.47%
2. 零-shot 分类：
    - Top-1 准确率：67.89%
    - Top-5 准确率：88.10%
3. 适配器微调：
    - Top-1 准确率：71.32%（多次实验的平均值）
    - Top-5 准确率：92.24%

四、实验结果分析

1. 添加适配器显著提高了模型性能，特别是在 Top-1 准确率上，表明通过微调模型可以更好地对齐图像和文本特征，从而提高分类准确性。
2. 与线性探针相比，适配器方法的效果更好，而零-shot 分类方法尽管未进行微调，但在没有微调的情况下仍能提供合理的性能，特别是在 Top-1 准确率上尚有提升空间。
3. 所有方法的 Top-5 准确率都较高，说明模型能在前几个预测中给出正确的类别，尽管 Top-1 准确率和 Top-5 之间的差距表明，模型能接近正确分类，但未必每次都能将最可能的类别排在第一。

五、结论

- 本实验展示了结合图像和文本特征进行多模态分类任务的有效性。
- 未来可以进一步优化适配器模型，或尝试其他架构来提升特征对齐与表示效果。此外，探索不同的数据集或增加训练数据可能进一步提升准确性。